FOCUS OF ATTENTION IN A DISTRIBUTED-LOGIC SPEECH UNDERSTANDING SYSTEM

Frederick Hayes-Roth and Victor R. Lesser
Computer Science Department[1]
Carnegie-Mellon University
Pittsburgh, Pa. 15213

January 12, 1976

DEPARTMENT of COMPUTER SCIENCE

Carnegie-Mellon University

FOCUS OF ATTENTION IN A DISTRIBUTED-LOGIC SPEECH UNDERSTANDING SYSTEM

Frederick Hayes-Roth and **Victor R. Lesser**
Computer Science Department[1]
Carnegie-Mellon University
Pittsburgh, Pa. 15213

January 12, 1976

ABSTRACT

The Hearsay II speech understanding system under development at Carnegie-Mellon University is a complex, distributed-logic processing system. Processing in the system is effected by independent, data-directed knowledge source processes which examine and alter values in a global data base representing hypothesized phones, phonemes, syllables, words, and phrases, as well as the hypothetical temporal and logical relationships among them. The question of how to schedule the numerous potential activities of the knowledge sources so as to understand the utterance in minimal time is called the "focus of attention problem". Near optimal focusing is especially important in a speech understanding system because of the very large solution space that potentially needs to be searched. Using the concepts of stimulus and response frames of scheduled knowledge source instantiations, competition among alternative responses, goals, and the desirability of a knowledge source instantiation, a general attentional control mechanism is developed. This general focusing mechanism facilitates the experimental evaluation of a variety of specific attentional control policies (such as best-first, bottom-up, and top-down search heuristics) and allows the modular addition of specialized heuristics for the speech understanding task.

1 This research was supported in part by the Defense Advanced Research Projects Agency under contract no. F44620-73-C-0074 and monitored by the Air Force Office of Scientific Research.

INTRODUCTION

The Hearsay II (HSII) speech understanding system (Lesser, et al., 1974; Erman & Lesser, 1975) is a complex, distributed-logic processing system. Inputs to the system are temporal sequences of sets of acoustic segments and associated hypothesized labels. Diverse sorts of speech understanding knowledge are encoded in several (15, currently) independent knowledge source modules (KSs), which include one or more KSs specific to each of the following knowledge domains: acoustic-phonetic mappings, phone expectation-realization relationships, syllable recognition, word hypothesization, and syntax and semantics. The state of processing at any point in time is represented by a global data base (the blackboard) which holds in an integrated manner all of the current hypothesized elements, including alternative guesses, at the various information levels of interpretation (e.g., segmental, phonetic, phonemic, syllabic, word, and phrasal). In addition, any inferred logical or confirmatory relationships among various hypotheses are represented on the blackboard by weighted and directed links between associated hypotheses. The weight and direction of a link reflect the degree to which the hypothesis at the tail of the link implies (supports or confirms) that at the head. The blackboard may be viewed as a two-dimensional problem space, where the time and information level of a blackboard hypothesis serve as its coordinates. Such a view permits consideration of specific "areas" of the problem space and enables us to speak meaningfully of hypotheses in the "vicinity" of a specific data pattern.

Processing in the system consists of additions, alterations, or deletions made to data on the blackboard by the various KSs. Each KS is data-directed, i.e., it monitors the blackboard for arrival of data matching its precondition pattern, a particular pattern of hypotheses and links and specific values of their attributes. Whenever its precondition is matched, the KS is invoked to operate separately on each satisfying data pattern. Finally, when the KS is executed, its (arbitrarily complex) logic is evaluated to determine how to modify the data base in the vicinity of the precondition pattern that triggered the invocation. The data pattern matching the precondition of a KS will be denoted as the stimulus frame (SF) of the invocation, and the changes it makes to the data base as its response frame (RF). Each KS may be schematized as a production rule of the form [precondition => response]. Each instantiation is then schematized [SF => RF], reflecting the fact that the RF data pattern is produced in response to the determination that the SF matches the rule's precondition. Because of the complexity of knowledge source processing, a precise definition of the RF cannot be directly calculated from the stimulus frame without the actual execution of the

knowledge source. However, an abstraction of the RF which specifies the type of changes that may be made (e.g., the addition of a new hypothesis or new link, the modification of a hypothesis' validity, etc.,) and the general vicinity of the changes can be easily calculated directly from the SF. It is this abstraction of the RF which will be used in further discussions.

As is well known in speech understanding research, each KS is imperfect. At any level of analysis, a very large number of errors may be introduced, including misclassifications, failures to recognize, and inappropriate "don't care" responses to what is actually a significant portion of the utterance. The common approach in speech understanding research is to construct systems which can recognize utterances in spite of such errors by evaluating many weakly supported alternative hypothesized interpretations of the speech simultaneously. A practical consequence of this parallel evaluation of numerous alternatives is that, at any point in time, a great number of KS applications are warranted by the existence of hypothetized interpretations matching the various KS preconditions. One object of attentional control is to schedule the numerous potential activities of the KSs to prevent the intractable combinatorial explosion which would inevitably result from an unconstrained application of KSs. More specifically, the focus of attention problem is defined to be that of developing a method for minimizing the total number of KS executions (or total processing time) necessary to achieve an arbitrarily low rate of error in the semantic interpretation of utterances.

The standard approach to the focus of attention problem in other speech systems employing diverse, cooperating sources (Reddy, et al., 1973; Paxton and Robinson, 1975; Woods, 1974) is based on an explicit control strategy. In these explicit control strategies, there is a centralized focusing module which carries out two functions using a built-in set of speech-specific rules: (1) for defining an explicit sequence of calls to a predefined set of knowledge sources and then evaluating their responses in order to determine the suitability of a hypothesized phrase (partial parse of the utterance); and (2) for deciding which of many alternative partial parses of the utterance should be further evaluated. This explicit control strategy is inappropriate in the HSII framework because it destroys the data-directed nature and modularity of knowledge source activity. In the HSII system, KSs can be easily removed or added, and their input and output characteristics changed without effecting other knowledge in the system. There is also a more fundamental argument against an explicit control strategy in a problem-solving system that uses a large number of diverse sources of knowledge: this explicit strategy requires the use of built-in knowledge about the

specific characteristics of knowledge sources. In this case, it seems that the explicit sequential logic necessary to get the appropriate interactions among the knowledge sources in all the possible different data patterns will become very difficult to predetermine and code.

The approach taken in HSII to focus of attention does not use any explicit (pre-compiled) information about which knowledge sources currently are contained in the system, nor their processing characteristics; this approach is more Implicit (i.e., mechanistic, uniform, and data-directed); it relies more on general task independent focusing strategies than on speech-specific ones. It should also be noted that, as part of these more general focusing strategies employed in HSII, a uniform mechanism has been incorporated which allows a knowledge source to contribute speech-specific focusing information through modifications to the blackboard. In this way, speech-specific focusing information can be exploited without destroying the modularity and the data-directed nature of knowledge source control in the HSII systems framework.

The remainder of this paper is divided into four sections. In the next section, a number of underlying principles for effective focusing and related processing control mechanisms are described. Subsequently, in the section on "Additional Mechanisms for Precise Focusing," additional objectives for focusing are discussed and related mechanisms for their attainment are presented. The section on "Alternative Policies for Focus of Attention" describes how these techniques permit experimentation with a variety of attentional control policies, such as purely bottom-up, purely top-down, and hybrid analyses. Finally, tentative conclusions are discussed in the last section.

FUNDAMENTAL PRINCIPLES AND MECHANISMS

One can view the focusing problem as a complex resource allocation problem. For example, consider the expenditure of money on alternative search devices in a hunt for oil. The alternative explorers and devices, including seismologists, geologists, drilling teams, and satellite reconnaissance, are the knowledge sources of the task. Each produces its response data only with significant cost and with a substantial probability of error, and there are sequencing constraints which require some KSs to delay their processing until other KSs terminate theirs and then only if particular findings are obtained. How should one invest in their potential contributions? Five fundamental principles have been identified for the control of processing in such tasks, and these are listed below. Each of these principles is used to define a separate

measure for evaluating the importance that should be attached to each KS invocation that has not yet been executed. These measures that are associated with each KS invocation are not necessarily constant for the lifetime of the invocation but may need to be dynamically recalculated as the state of the blackboard changes in the general vicinity of KS's stimulus and response frame. A function based on these measures is then used to associate a priority to each KS invocation.

(1) The competition principle: the best of several alternatives should be performed first. This principle governs how ordering decisions should be made among several behavioral options which are competitive in the sense that a successful outcome of one obviates performing another. For example, consider the problem of determining whether oil exists at site A and suppose that the functions of a geologist and seismologist are substitutable vis-a-vis this objective. If either the seismologist or geologist has already performed and positively indicated the presence or absence of oil, that result obviates employing the other scientist to perform an equivalent function. In this sense, it can be said that the previous result competes with the yet-to-be-performed alternative; that is, the former response is at a higher level of analysis in the same area of the problem space as is the alternative action. However, if oil on site B can be determined only by seismological techniques, hiring a geologist for site A does not compete with hiring a seismologist for site B, according to this principle.

(2) The validity principle: more processing should be given to KSs operating on more valid data. This principle says that, everything else constant, one KS invocation should be preferred to another if the former is working on data which is more credible. In an oil hunt, it would be preferred to employ as a predictor the one seismologist whose seismological readings were most accurate. Similarly, in the speech domain, various KSs will be invoked to contribute to the interpretation of specific data patterns on the blackboard. Each hypothesis in a SF will contain a rating of its validity derived from the validities and implications of hypotheses linked to it. Thus, this principle implies that the KSs invoked to work on the most valid SFs are most preferred. Once these KSs have performed, the hypotheses in their responses will also be rated for validity and will, in general, derive their validity directly from the hypotheses in the SF. By preferring KS invocations with the most credible SFs, the system tends to maximize the validity of its responses.

(3) The significance principle: more processing should be given to KSs whose RFs are more significant. This principle aims at insuring that when a variety of behaviors can be performed, the most important are done first. For example, while

filing a claim on land and drilling are both necessary prerequisites for successful completion of an oil hunt, at the outset of prospecting the former is the more important and should be done first. As an example in the speech domain, a situation might arise where a sequence of phones could be either recognized as a word or subjected to analysis for coarticulation effects. The first of these two actions is more important and, on a priori terms, should be performed first. One heuristic in the speech understanding domain for defining significance is to give preference to KS invocations which are operating at the highest levels of analysis within any portion of the utterance (closest to a complete parse interpretation). A more general statement of this heuristic is that preference should be given to the KS invocation whose RF can potentially produce a result which is closest in terms of information level to the overall goal of the problem solver.

(4) The efficiency principle: more processing should be given to KSs which perform most reliably and inexpensively. Obviously, if one geologist is more reliable than another and the two charge the same for their services, the former should be preferred. Conversely, of two equally reliable geologists, one should prefer the less expensive. Similarly, in the speech domain, many KS applications are more efficient than others and should be preferred. As an example, a bottom-up word hypothesizer is found to be more accurate at generating word hypotheses than is the top-down syntax and semantics KS. Everything else equal, two invocations of these KSs whose response frames consist of new word hypotheses should be scheduled so that the bottom-up hypothesizer is first executed.

(5) The goal satisfaction principle: more processing should be given to KSs whose responses are most likely to satisfy processing goals. The oil hunt managers might establish a goal of determining the depth of water at site A. This would induce additional preference for those agents (e.g., the seismologists and drillers) whose ordinary activities could concomitantly satisfy this additional goal. In the speech domain, similar circumstances arise: the priority of a KS which can potentially generate new word hypotheses in a particular time region of the utterance should be increased. This desire for a specific type of processing is specified in HSII by establishing a goal on the blackboard which represents the time and level of the desired hypotheses. KS instantiations whose RFs match the processing specified in the goal are made more desirable. More generally, KS invocations may be evaluated as more or less likely to help satisfy each specific goal. The higher the probability that a KS invocation will contribute to the satisfaction of a goal and the greater the utility of the goal, the more desirable its execution becomes. Through this mechanism of adding

goals to the blackboard, a knowledge source can dynamically introduce task specific focusing rules into the focusing algorithm. Since KS activity is data-directed, this focusing policy KS would execute only when the data patterns indicating the need for a specific focus action occur.

The preceding five principles provide the theoretical foundation for our attentional control system. A number of sophisticated control mechanisms have been created which provide the tools by which these principles can be converted into operational focusing policies. These mechanisms are discussed in the remainder of this section.

In order to evaluate the preferability of one KS invocation vis-a-vis the others, the five control principles require a number of ordering relationships to hold. In overview, the major operational principle for focusing is to schedule for earliest execution the KS invocation which is the most desirable according to the five rules provided. The focusing mechanism first evaluates the desirability of each KS invocation as a measure of the degree to which it satisfies the various objectives of the system and then executes the most desirable first (with an appropriate generalization for executing several KSs simultaneously in a multiprocessing system). Thus, the major subproblem in the construction of a focuser is the estimation of a KS invocation's desirability. How this desirability is computed will now be described.

Each KS invocation is characterized by a number of attributes. Its SF has a credibility value (between -100 and +100) which estimates the likelihood that the detected pattern of hypotheses and links is valid and satisfies the KS's precondition (negative values imply evidence against this possibility). The credibility value of a SF is determined as a function of the validity ratings on each of the hypotheses in the SF. As previously indicated, these ratings themselves are determined from the strengths of implications on links, the original probabilities assigned to each of the acoustic segment labels provided as input (i.e., the lowest level hypotheses in the blackboard), and the derived validity ratings of intermediate level hypotheses. In our current implementation, the credibility of the SF is taken to be the maximum of the validity ratings of the hypotheses in the SF (ranging from -100 to +100).

Each KS invocation can be thought of as a transformation of the SF into the RF. Associated with the KS invocation then is the estimated level(s) (e.g., phonetic, word, phrasal) of the RF, the estimated validity of the RF hypotheses, and the estimated time (i.e., location and duration) of any newly created RF hypotheses. Each of these estimated values contributes to an appraisal of the ignificance and probable correctness of the RF which the KS will produce.

The objectives of the significance, efficiency, and goal satisfaction principles can be achieved if the desirability of a KS invocation is computed by any increasing function of the credibility of its SF, the estimated reliability of the KS (to produce correct RFs of the form it anticipates), and the estimated level, duration, and validity of RF hypotheses. The objective of the validity principle, to operate on most valid data first, is accomplished by making desirability an increasing function of the credibility of the SF. The objective of the significance principle, to perform the most significant behaviors first, is achieved by making desirability an increasing function of the level and duration of RF hypotheses. Since hypotheses closest to complete utterance interpretations will be at the highest level and span the entire duration of the speech, actions which can produce such hypotheses or support them will be most preferred. The objective of the efficiency principle, to prefer KSs which perform best, is achieved by making desirability an increasing function of the KSs reliability (per unit "cost" or time).

To understand how the other objectives, the preference of the competition principle for avoiding computation of obviated behaviors and the goal-directed scheduling dictated by the goal satisfaction principle, are achieved in the system, it is necessary to introduce a number of additional concepts. The mechanisms required to operationalize the desired effects of competition will be considered first.

The first objective of the focuser is to insure that the understanding system moves quickly to a complete interpretation of the speech and, in particular, avoids apparently unnecessary computation. Specifically, if any KS invocation is expected to produce a RF which is in the same time range as an existing, higher level, longer duration, and more credible hypothesis, its activity is potentially useless. It is therefore less preferred than the action of a KS which is expected to produce higher level, more expansive, and more credible interpretations of the utterance than those that currently exist. Thus, HSII uses a statistic called the state of the blackboard; this is a single-valued function of each time value, from the beginning of an utterance to its end. The state S(t) for some point (time) t in the utterance is the maximum of the values V(h) of all hypotheses which represent interpretations containing the point t. The value of a hypothesis is an increasing function of its level, duration, and validity. Thus, the highest possible value for a hypothesis would be that associated with the hypothesis representing a complete parse of the entire utterance with a validity rating of +100 (the maximum). To the extent that the utterance is partially parsed in some interval [t1,t2], will the state S(t) be high in this region. Thus, S(t) provides a single metric for evaluating the current success of the understanding process over each area

of the utterance. From a more general viewpoint, the metric V(h) indicates how close a hypothesis h is to the desired overall goal state; and, the metric S measures both what aspect of the overall goal has been solved (e.g., in the case of speech, what time interval) and how good is the solution (e.g., in the case of speech, the validity of the hypothesis and how close in terms of information level it is to the sentential phrase).

It is very easy, using S(t), to decide whether a prospective action is likely to improve on the current state of understanding. If the estimated value V(h) of a RF hypothesis h exceeds S(t) anywhere in the corresponding interval, the KS invocation should be considered very desirable; otherwise it should be inhibited by the existing more valuable, competitive hypotheses. This, in short, is how the objective of the competition principle is accomplished. In addition to its dependence upon the variables already considered, the desirability of a KS invocation is made to be an increasing function of the ratio of the maximum of the estimated value of the RF hypotheses to the current state S(t) (where S(t) is taken to be the minimum over the interval corresponding to the time location of the RF). In this way, preference is given to KS invocations which are expected to improve the current state of understanding.

One can think of S(t) as defining a surface whose height reflects the degree of problem solution in each area. In this conception, operations which would yield results below the surface are undesirable (unnecessary), and those which would raise the surface are preferred.

The last objective to be operationalized is that of the goal satisfaction principle. In general, a goal may specify that particular types of hypotheses are to be created (e.g., create word hypotheses between times t_0 and t_1) or existing hypotheses modified in desired ways (e.g., attempt to reject the hypothesized word "no" between t_3 and t_4 by establishing disconfirming relationships between it and the acoustic data). Two types of adjustments are made to the desirability ratings of KS invocations based on their relationships to such goals. The first case arises when there is direct goal satisfaction, meaning that a KS invocation is a possible candidate for solving a goal because its RF matches the desired attributes of the goal. In this case, the desirability of the KS invocation is increased by an amount proportional to the utility of the goal (the degree to which it is held to be important when it is created).

The second type of effect is the result of indirect goal satisfaction. In this case, a KS invocation does not directly satisfy a goal but apparently increases the probability that it will be solved by producing some result which is held to be partially useful for the achievement of the main goal. Two types of indirect goal satisfying actions can be identified. First, there is goal reduction: a KS invocation generates

subgoals whose solution(s) will entail satisfaction of the original goal. For example, as the result of recognizing the sequence "The (gap) dog," the system might establish a goal for the recognition of an adjective between the two recognized words to replace the gap in understanding. Subsequently, some KS might establish several disjunctive subgoals related to this one, such as goals for recognizing the words "shaggy," "cute," "sleepy," etc. Because the satisfaction of any one of these would constitute satisfaction of the original objective, the KS invocation indirectly satisfies the original goal. Its desirability is less than that of a KS invocation directly satisfying the same goal, but may be more than other KSs.

The second type of indirect goal satisfaction occurs when a KS invocation approaches a goal by producing a RF which is close to the goal but does not quite satisfy it. For example, in the context of the preceding "adjective" goal, a general increase in the activity of knowledge sources which generate and improve phone hypotheses, syllable hypotheses, and phrasal hypotheses in the area of interest will be more or less proximate to the desired response. Since each KS is schematized as a rule of the form [precondition => response], a means-ends analysis can be performed to estimate the probability that some KS invocation will produce a response contributing to the ultimate solution of a goal. The more closely its RF approaches the desired goal, the higher is the probability that execution of a KS invocation will contribute to the goal's ultimate satisfaction and the greater the desirability of the KS invocation.

In summary, the desirability of a KS invocation is defined to be an increasing function of the following variables: the estimated value of its RF (an increasing function of the reliability of the KS and the estimated level, duration, and validity credibility of the hypotheses to be created or supported); the ratio of the estimated RF value to the minimum current state in the time region of the RF; and, the probability that the KS invocation will directly satisfy or indirectly contribute to the satisfaction of a goal as well as the utility of the potentially satisfied goal. Scheduling KS invocations according to their desirabilities then accomplishes the objectives established by the preceding five basic principles. However, there are some inadequacies of such a basic attentional control mechanism; these are considered in the next section.

ADDITIONAL MECHANISMS FOR PRECISE FOCUSING

Basically, while the five fundamental principles appear correct and universally

applicable, they are not complex enough to provide precise control in all of the situations that arise in a complex distributed-logic understanding system. Three additional issues are now introduced, and the control mechanisms currently used to handle these are discussed. The topics considered include dynamically modifiable recognition and output generation thresholds on KS logic; an implicit goal state (approximately the inverse of the current state S(t)) which can be used to determine the desired balance between depth-first and breadth-first approaches to the understanding problem; and methods for avoiding "false peaks" or "cognitive fixedness" in the recognition process.

Nearly all KS behavior can be separated into two components: a pattern recognition component and an output generation component. For example, a word hypothesizer may look for patterns of phones (pattern recognition) in order to produce a new word hypothesis (output generation). Both components operate in fuzzy, errorful ways. In the pattern recognition component, the KS must accept fuzzy matches of its templates because that is the nature of speech recognition. Conversely, the word hypotheses it generates are necessarily probabilistic. The probable correctness of its hypotheses are then reflected by validity ratings or implication weights on its outputs. Thresholding occurs in such processes in two ways. First, the degree of fuzziness tolerated in pattern matching is arbitrarily set to some moderate criterion to prevent an intractably large number of apparent matches. Second, the strengths of the output responses are measured against some threshold to insure that only sufficiently credible responses are produced. The credibility of the response may, in addition to its dependence upon the credibility of the stimulus frame, also be dependent upon the type of inference method used to generate a response. For example, the word recognizer might employ a distance metric for recognition and classification, in which case the credibility of the output word is a decreasing function of the distance between the stimulus phones and the phones of the most similar word template. Responses which are too weak vis-a-vis this second threshold are held in abeyance rather than being produced or forgotten.

Now the general scheme of the robust overall policy that is employed can be sketched. At the beginning of an analysis, relatively high thresholds are specified for pattern matching goodness and output goodness. Processing continues based on the other scheduling principles until thresholds are changed (discussed below). When a threshold change occurs, it may be specific to certain levels or time regions of RFs or to the types of KSs used to produce them. As an example, if all of the utterance were correctly understood except the first word, we would set very low thresholds for

behavior for all KSs in the beginning portion of the utterance. Our current policy, in specific, lowers thresholds most in poorly understood areas adjacent to areas which are well understood. When an arbitrary level of desirability is no longer achieved by any of the pending KS invocations, the important areas for threshold lowering are identified by finding valleys next to peaks in the state function S(t). The thresholds in these areas are lowered in the hope that greater error tolerance there will produce additional results which can be usefully integrated with the adjacent, more reliable interpretations previously produced.

Without dynamically modifiable pattern match and output goodness thresholds, a speech understanding system would necessarily embody numerous parameters whose values were determined at the outset for all problem tasks. Such a system would probably be very sensitive to the particular values chosen. Our approach, however, insures that each of the KSs can be encouraged to perform more work in any area of the blackboard by simply lowering two general sorts of control variables. This is seen as a fundamentally important control principle relating to the controllability of the generative aspect of KSs _per se_ rather than to their comparative expected responses.

The second additional concept which is utilized in the focuser is that of the _implicit goal state_ or I(t). It is only a slight oversimplification to think of I(t) as the inverse of the current state S(t). To the extent that S(t) is large (representing the fact that the portion of the utterance adjacent to t has been highly successfully analyzed), I(t) will be small. A small I(t) value means that there is little to be gained by trying to improve the understanding around t. Conversely, a large I(t) means that the portion of the utterance in the neighborhood of t greatly needs additional analysis. As a result, one might suppose that KSs operating in that region should be conceived as satisfying an implicit goal of raising the level of understanding (the surface of the current state S(t)) wherever it is lowest. In fact, the best role for the implicit goal state is probably as a weak contributor to the desirability of a KS invocation. It remains an empirical question whether it is better to work in the regions of the highest peaks in understanding (_depth-first_) or more evenly throughout the entire utterance (_breadth-first_). Although an optimal strategy is not known, it is clear that in computing the desirability of a KS invocation, the estimated value of the RF and the ratio of the RF value to the minimum of S(t) in the same region are two contributing factors whose relative weightings can be experimentally manipulated to achieve exactly that balance between depth-first and breadth-first which is desired.

As is well known in problem solving and search paradigms, there is a constant danger of getting trapped on "false peaks," as when one bases actions on the apparent

correctness of highly rated but ultimately incorrect interpretations. A number of the preceding focusing principles have been formulated to insure that processing in the region of highly valued hypotheses is facilitated at the expense of other potential actions; a consequence of this paradigm is that the focuser must take precautions to prevent the "cognitive fixedness" which would be apparent if the focuser failed to abandon those paths which lead nowhere. This is done in the focuser in a simple manner. The highest peak in understanding at any point t in the utterance corresponds to the highest valued hypothesis in that region, and its value is just S(t). Thus, stagnation of the understanding process in a region can be detected whenever S(t) fails to increase for a prolonged time. While preference should still be given to the execution of KS invocations working on the surface of S(t) and promising to increase its value, the focuser must conclude that other KS invocations should now become more desirable than they previously seemed, because they at least may improve the analysis in the stagnant area. This is accomplished by increasing the implicit goal state I(t) whenever S(t) is stagnant for a specified length of time. As a result of increasing I(t), KS invocations operating near the surface of S(t) and previously viewed as marginally desirable become sufficiently desirable to be executed. If any one of them succeeds in increasing S(t), I(t) is promptly reset to be the inverse of S(t). However, each time S(t) stagnates for the specified duration, I(t) is again increased. Thus, false peaks are avoided by actually recognizing the behavioral characteristics of cognitive fixedness: as long as the degree of its understanding remains stagnant, it continually increases the desirability of the competing KS alternatives which previously appeared to be suboptimal in the area of stagnation.

ALTERNATIVE POLICIES FOR FOCUS OF ATTENTION

To this point, general principles for focusing and mechanisms to achieve the realization of these principles have been described. However, there still remains a wide variety of policies which can be superimposed upon these mechanisms in a manner consistent with them but prescribing a specific global search strategy to be employed in speech understanding. This flexibility is considered one of the outstanding virtues of the focuser design since it affords the possibility for empirical evaluation of alternative focus of attention policies. In this section, a number of these policies are identified, and it is shown how each of these can be easily effected within our system. Each policy described would be effected by one or more policy modules, a

KS-like program which is activated whenever specific conditions of interest are detected. This will be clarified by the examples below.

Consider the policy which dictates that, whenever possible, understanding is to proceed bottom-up, from the acoustic segments to the phrasal level. Such a policy would be effected as follows. At the outset the policy module would set a goal with infinite positive utility for RFs at the lowest level and a goal with infinite negative utility for RFs at higher levels. When the system became quiescent, the policy module would be reinvoked by the system. Its response would be to modify the goals so that processing at the two lowest levels would be facilitated and all others inhibited. This process would continue until the highest level was facilitated. At any particular point in the analysis, processing would be restricted to several of the lowest levels and would move upward one level at a time as all the potential activity at a lower level had been completed. Similarily, a purely top-down analysis could be controlled in the same way, substituting "highest" for "lowest", etc.

Under ordinary circumstances, using only the mechanisms detailed in the previous sections, a hybrid analysis will occur. While there is increased desirability associated with RFs at the highest levels, it is to be expected that sometimes there will be areas of the utterance where all desirable KS invocations will be at low levels while in other areas they will be primarily at higher levels.

A left-to-right analysis can be accomplished using goals in the same way as for the purely bottom-up or top-down methods. Here, every time quiescence occurs, the processing from the beginning of the utterance to a point further along in time is facilitated. This would continue until the whole utterance was facilitated by a goal. Right-to-left, obviously, is similarly controlled. Note too that "more or less" left-to-right search can be accomplished by specifying less than infinite goal utilities and by defining "quiescence" to mean that the desirabilities of all KS invocations are below some policy threshold for minimally acceptable desirability.

Perhaps one of the most important types of empirical comparisons to be studied is the breadth vs. depth-first alternatives. Breadth-first is, thecretically speaking, advantageous when KSs are capable of looking at broad contexts and optimizing their outputs on the basis of more information than is used, for example, by simple grammatical rewriting rules. Similarly, if KSs are capable of appreciating the extent to which various hypotheses are partially supported by disparate but cooperative data scattered about the blackboard, a breadth-first approach should exhibit some "intelligence". Alternatively, a depth-first approach is desirable whenever KSs make few errors. For example, if word recognition becomes very good, then it should be

possible to rely upon the words and upon the inferences (e.g., other predicted words) which are derived from them. This reduction in the necessary parallelism of hypothesization makes depth-first a reasonable strategy. In the interim, however, it is apparent that there may be enormous differences in the overall system performance under these different control policies. It is hoped that in the near future empirical data on the relative utility of these different strategies can be obtained. Moreover, if the relative effectiveness of these different control strategies can be associated with formal properties of a problem's structure and complexity, it may be reasonable to anticipate that such empirical observations will be helpful in evaluating the formal complexity of the speech understanding problem.

In summary, it is suggested that the principles and mechanisms described in the preceding sections provide a parameterized framework for the elaboration of numerous alternative "macroscopic" policies for attentional control in the speech understanding problem. Each of the typical sorts of heuristic problem solving policies can be realized by simple policy modules which manipulate goal utilities and respond to quiescence in policy-specific ways.

SUMMARY

By schematizing knowledge sources as [precondition => response] rules, each potential behavior of the Hearsay II system is viewed as an instantiation of such a form. These KS instantiations are seen to be [stimulus frame => response frame] action descriptions. The desirability of an instantiation is then computable from several characteristics of the stimulus and response frames. By enumerating the fundamental principles for attentional control, a desirability measure is produced which handles most of the problems in focusing. Several additional objectives make elaboration of this simple strategy desirable. In order to accomplish more precise overall control, computations are made of the current state of the analysis, the implicit goal state of the system, and the relative degree of goal satisfaction of each KS invocation. Once the desirability of each KS invocation is computed, the execution of the most desirable first serves to accomplish an apparently optimal allocation of computing resources. In addition, our framework provides an excellent environment in which to explore empirically the utility of many global focusing strategies. Each of these can be expressed in terms of particular weightings of the contributions of various terms to the desirability of a KS invocation or by simple modules which create,

modify, and monitor goals which control the direction of analysis. The relatively small grain size of knowledge representation and fine identification of the type and location of knowledge source contributions apparently affords great advantages in constructing mechanisms to control a large, distributed, knowledge-based understanding system.

ACKNOWLEDGMENTS

We would like to acknowledge the help of the following people in the design and implementation of these ideas in the HSII system: Donald Kosy, Craig Everhart, and David McKeown.

REFERENCES

Erman, L. D., & Lesser, V. R. A multi-level organization for problem solving using many, diverse, cooperating sources of knowledge. Proc. of the 4th IJCAI, 1975.

Lesser, V. R., Fennel, R. D., Erman, L. D., & Reddy, D. R. Organization of the HEARSAY II speech understanding system. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1975, ASSP-23, 11-23.

Paxton, W. H., & Robinson, A. E. System integration and control in a speech understanding system. A. I. Center, Tech. Note 111, SRI, Menlo Park, Ca. 1975.

Reddy, D. R., Erman, L. D. & Neely, R. D. A model and a system for machine recognition of speech. IEEE Trans. Audio and Electroacoustics AU-21,3, 1973, 229-239.

Woods, W.A. Motivation and overview of BBN SPEECHLIS: an experimental prototype for speech understanding research. Proc. of IEEE Symposium on Speech Recognition, Carnegie-Mellon Univ., Pittsburgh, Pa., 1974, 1-10.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR - TR - 76 - 0596	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): FOCUS OF ATTENTION IN A DISTRIBUTED-LOGIC SPEECH UNDERSTANDING SYSTEM	5. TYPE OF REPORT & PERIOD COVERED: Interim report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Frederick Hayes-Roth, Victor R. Lesser	8. CONTRACT OR GRANT NUMBER(s): F44620-73-C-0074, ARPA Order-2466
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Carnegie-Mellon University, Computer Science Dept., Pittsburgh, PA 15213	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61101D, AO 2466
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: Jan[redacted]976
	13. NUMBER OF PAGES: 16
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Air Force Office of Scientific Research (NM), Bolling AFB, DC 20332	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

see back side of page

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

The Hearsay II speech understanding system under development at Carnegie-Mellon University is a complex, distributed-logic processing system. Processing in the system is effected by independent, data-directed knowledge source processes which examine and alter values in a global data base representing hypothesized phones, phonemes, syllables, words, and phrases, as well as the hypothetical temporal and logical relationships among them. The question of how to schedule the numerous potential activities of the knowledge sources so as to understand the utterance in minimal time is called the "focus of attention problem". Near optimal focusing is especially important in a speech understanding system because of the very large solution space that potentially needs to be searched. Using the concepts of stimulus and response frames of scheduled knowledge source instantiations, competition among alternative responses, goals, and the desirability of a knowledge source instantiation, a general attentional control mechanism is developed. This general focusing mechanism facilitates the experimental evaluation of a variety of specific attentional control policies (such as best-first, bottom-up, and top-down search heuristics) and allows the modular addition of specialized heuristics for the speech understanding task.